

Atty. Docket No. MS149546.02

CLUSTER-BASED VISUALIZATION OF USER
TRAFFIC ON AN INTERNET SITE

by

Igor Cadez, David E. Heckerman, Christopher A. Meek,
and Steven J. White

MAIL CERTIFICATION

I hereby certify that the attached patent application (along with any other paper referred to as being attached or enclosed) is being deposited with the United States Postal Service on this date April 22, 2004, in an envelope as "Express Mail Post Office to Addressee" Mailing Label Number EV373132592US addressed to the Mail Stop Provisional Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450.


Deborah L. Corpus

Title: CLUSTER-BASED VISUALIZATION OF USER TRAFFIC ON AN
INTERNET SITE

5 CROSS-REFERENCE TO RELATED APPLICATIONS

This application is a continuation of U.S. Patent Application Serial No. 09/517,462, filed on March 2, 2000, entitled "CLUSTER-BASED VISUALIZATION OF USER TRAFFIC ON AN INTERNET SITE", the entirety of which is incorporated herein by reference.

10 FIELD OF THE INVENTION

This invention relates generally to visualizing traffic on Internet world-wide-web sites, and more particularly to such visualization that is cluster-based in nature.

BACKGROUND OF THE INVENTION

15 Browsing the Internet has become an exceedingly popular application for computers and computer-like devices. In particular, sites on the world-wide-web (WWW, or web) have seen increasing visitation. A web site includes a number of web pages, identified by Uniform Resource Locator (URL) addresses. For example, a home page of a web site may be addressed as <http://www.mypage.com/index.html>.

20 As web sites have become more popular, tracking their usage has becoming important, especially for commercial sites. Operators of web sites, for example, are interested in determining how users proceed through the site in a single session. A session is when a user views one or more pages of the web site, in a particular order, before going to a different web site. By being able to determine how users are proceeding
25 through their sites, operators are able to better understand their users, and how to make their web sites better.

For this and other reasons, therefore, there is a need for the present invention.

SUMMARY OF THE INVENTION

30 The invention relates to visualizing web traffic in a cluster-based manner. In one embodiment, a number of windows are displayed, corresponding to a number of clusters into which users have been partitioned based on similar web browsing behavior. The

windows are ordered from the cluster having the greatest number of users to the cluster having the least number of users. Each window has one or more rows, where each row corresponds to a user within the cluster. Each row has an ordered number of visible units, such as blocks, where each block corresponds to a web page visited by the user.

- 5 The blocks can be color coded by the type of web page they represent. In one embodiment, the corresponding cluster models for the clusters are alternatively displayed in the windows. The windows desirably can be made larger or smaller, as well as be scrolled through via scroll bars.

Thus, embodiments of the invention allow web site operators to view first-hand
10 how users are visiting their web sites. The visual approach of embodiments of the invention provides for advantages as compared to non-visual approaches, since usage data for web sites can be increasingly large. Without a visual approach, for instance, it is more difficult for operators to make conclusions as to usage of their web sites, as compared to using a visual approach.

15 The invention includes computer-implemented methods, machine-readable media, computerized systems, and computers of varying scopes. Other aspects, embodiments and advantages of the invention, beyond those described here, will become apparent by reading the detailed description and with reference to the drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

20 FIG. 1 is a diagram of an operating environment in conjunction with which embodiments of the invention can be practiced;

FIG. 2 is a diagram of a display device on which a number of windows can be shown in accordance with an embodiment of the invention;

FIG. 3 is a diagram of a representative such window in accordance with an embodiment of the invention;

FIG. 4 is a diagram of a window showing a corresponding model for a cluster, according to an embodiment of the invention;

5 FIG. 5 is a flowchart of a method according to an embodiment of the invention; and,

FIG 6 is a diagram of a system according to an embodiment of the invention.

DETAILED DESCRIPTION OF THE INVENTION

In the following detailed description of exemplary embodiments of the invention,
10 reference is made to the accompanying drawings which form a part hereof, and in which
is shown by way of illustration specific exemplary embodiments in which the invention
may be practiced. These embodiments are described in sufficient detail to enable those
skilled in the art to practice the invention, and it is to be understood that other
embodiments may be utilized and that logical, mechanical, electrical and other changes
15 may be made without departing from the spirit or scope of the present invention. The
following detailed description is, therefore, not to be taken in a limiting sense, and the
scope of the present invention is defined only by the appended claims.

Some portions of the detailed descriptions which follow are presented in terms of
algorithms and symbolic representations of operations on data bits within a computer
20 memory. These algorithmic descriptions and representations are the means used by those
skilled in the data processing arts to most effectively convey the substance of their work
to others skilled in the art. An algorithm is here, and generally, conceived to be a self-
consistent sequence of steps leading to a desired result. The steps are those requiring

physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical or magnetic signals capable of being stored, transferred, combined, compared, and otherwise manipulated.

It has proven convenient at times, principally for reasons of common usage, to
5 refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like. It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussions, it is appreciated that throughout the present invention, discussions
10 utilizing terms such as processing or computing or calculating or determining or displaying or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system's registers and memories into other data similarly represented as physical quantities within the computer system
15 memories or registers or other such information storage, transmission or display devices.

Operating Environment

Referring to FIG. 1, a diagram of the hardware and operating environment in conjunction with which embodiments of the invention may be practiced is shown. The description of FIG. 1 is intended to provide a brief, general description of suitable
20 computer hardware and a suitable computing environment in conjunction with which the invention may be implemented. Although not required, the invention is described in the general context of computer-executable instructions, such as program modules, being executed by a computer, such as a personal computer. Generally, program modules

include routines, programs, objects, components, data structures, etc., that perform particular tasks or implement particular abstract data types.

Moreover, those skilled in the art will appreciate that the invention may be practiced with other computer system configurations, including hand-held devices, multiprocessor systems, microprocessor-based or programmable consumer electronics, network PC's, minicomputers, mainframe computers, and the like. The invention may also be practiced in distributed computing environments where tasks are performed by remote processing devices that are linked through a communications network. In a distributed computing environment, program modules may be located in both local and remote memory storage devices.

The exemplary hardware and operating environment of FIG. 1 for implementing the invention includes a general purpose computing device in the form of a computer, including a processing unit 21, a system memory 22, and a system bus 23 that operatively couples various system components include the system memory to the processing unit 21. There may be only one or there may be more than one processing unit 21, such that the processor of computer comprises a single central-processing unit (CPU), or a plurality of processing units, commonly referred to as a parallel processing environment. The computer may be a conventional computer, a distributed computer, or any other type of computer; the invention is not so limited.

The system bus 23 may be any of several types of bus structures including a memory bus or memory controller, a peripheral bus, and a local bus using any of a variety of bus architectures. The system memory may also be referred to as simply the memory, and includes read only memory (ROM) 24 and random access memory (RAM).

A basic input/output system (BIOS) 26, containing the basic routines that help to transfer information between elements within the computer, such as during start-up, is stored in ROM 24. The computer further includes a hard disk drive 27 for reading from and writing to a hard disk, not shown, a magnetic disk drive 28 for reading from or writing to a removable magnetic disk 29, and an optical disk drive 30 for reading from or writing to a removable optical disk 31 such as a CD ROM or other optical media.

The hard disk drive 27, magnetic disk drive 28, and optical disk drive 30 are connected to the system bus 23 by a hard disk drive interface 32, a magnetic disk drive interface 33, and an optical disk drive interface 34, respectively. The drives and their associated computer-readable media provide nonvolatile storage of computer-readable instructions, data structures, program modules and other data for the computer. It should be appreciated by those skilled in the art that any type of computer-readable media which can store data that is accessible by a computer, such as magnetic cassettes, flash memory cards, digital video disks, Bernoulli cartridges, random access memories (RAMs), read only memories (ROMs), and the like, may be used in the exemplary operating environment.

A number of program modules may be stored on the hard disk, magnetic disk 29, optical disk 31, ROM 24, or RAM, including an operating system 35, one or more application programs 36, other program modules 37, and program data 38. A user may enter commands and information into the personal computer through input devices such as a keyboard 40 and pointing device 42. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 21 through a serial port interface 46

that is coupled to the system bus, but may be connected by other interfaces, such as a parallel port, game port, or a universal serial bus (USB). A monitor 47 or other type of display device is also connected to the system bus 23 via an interface, such as a video adapter 48. In addition to the monitor, computers typically include other peripheral
5 output devices (not shown), such as speakers and printers.

The computer may operate in a networked environment using logical connections to one or more remote computers, such as remote computer 49. These logical connections are achieved by a communication device coupled to or a part of the computer; the invention is not limited to a particular type of communications device. The
10 remote computer 49 may be another computer, a server, a router, a network PC, a client, a peer device or other common network node, and typically includes many or all of the elements described above relative to the computer, although only a memory storage device has been illustrated in FIG. 1. The logical connections depicted in FIG. 1 include a local-area network (LAN) 51 and a wide-area network (WAN) 52. Such networking
15 environments are commonplace in office networks, enterprise-wide computer networks, intranets and the Internet, which are all types of networks.

When used in a LAN-networking environment, the computer is connected to the local network 51 through a network interface or adapter 53, which is one type of communications device. When used in a WAN-networking environment, the computer
20 typically includes a modem 54, a type of communications device, or any other type of communications device for establishing communications over the wide area network 52, such as the Internet. The modem 54, which may be internal or external, is connected to the system bus 23 via the serial port interface 46. In a networked environment, program

modules depicted relative to the personal computer, or portions thereof, may be stored in the remote memory storage device. It is appreciated that the network connections shown are exemplary and other means of and communications devices for establishing a communications link between the computers may be used.

5 Cluster-based Web Traffic Visualization Techniques

In this section of the detailed description, various approaches and techniques for visualizing web site activity in a cluster-based manner are described. In the diagram of FIG. 2, for instance, on the display device 200 a series of windows 202a, 202b, 202c, . . . , 202d, 202m and 202n is shown. Each of the series of windows corresponds to a cluster
10 into which users of a given web site or web sites have been partitioned based on similar web browsing behavior through the site(s). A cluster is generally defined in a non-restrictive and a non-limiting manner herein as a group of like items that have been associated with one another based on one or more common or similar characteristics via a predetermined technique. For example, in one embodiment of the invention, an
15 Expectation Maximization (EM) approach is used, as known within the art, to yield a mixture model of first-order Markov models, as also known within the art, where each cluster has a corresponding first-order Markov model. In another embodiment, an Expectation Maximization (EM) approach is used, as known within the art, to yield a mixture model of zero-order Markov models, as also known within the art, where each
20 cluster has a corresponding zero-order Markov model. However, the invention is not so limited. The windows are desirably ordered from the cluster having the greatest number of users therein to the cluster having the least number of users therein. Each of the windows can desirably be increased in size, also known as maximized, so that more rows

of the window is shown on the device 200. Furthermore, desirably each of the windows can have scroll bars added thereto, as known within the art, so that the data within the window is navigable via the bars.

In the diagram of FIG. 3, a representative window 300, that could be any of the series of windows, is shown. The window has at least one row, such as the rows 302a, 302b, 302c, 302d, . . . , 302n. Each row corresponds to a particular session of a particular user, where the sessions of the users in a given cluster are grouped together based on similar web browsing behavior. Each row has an ordered number of visible units, such as, in one embodiment, blocks, although the invention is not so limited. As an example, the row 302a has ordered blocks 304a, 304b, 304c and 304d. Each of the ordered visible units of a row represents a page of the web site visited by the user of the row. For example, the row 302a indicates that in this session, a user has visited four pages of a site.

Desirably, the visible units are coded, such as by color, to indicate the different types of pages on the site. The types of web pages are referred to herein synonymously as the different parts of the site that can be visited by the user. Thus, for example, for a news-related web site, the visible units may be color coded as to the particular parts of the site, such as local news encompassing local news-related pages, global news encompassing global news-related pages, financial news encompassing financial news-related pages, sports news encompassing sports news-related pages, entertainment news encompassing entertainment news-related pages, etc. Thus, the path of a given user session within a given cluster through a web site can easily be determined via examination of the number and order of the visible units, as well as their color. Furthermore, in one embodiment, the rows within a window can be ordered based on a

predetermined typicality measure of the users to which the rows correspond. In one embodiment, the predetermined measure is the probability of behavior based on location within a cluster, or, $p(\text{behavior}|\text{cluster})$.

As has been described, in one embodiment, each cluster to which a window
5 corresponds has a corresponding first-order Markov model. Therefore, in one
embodiment, a selected window or windows, or all the windows, can be made to display
the corresponding Markov model(s) for these window(s), in lieu of the ordered rows
corresponding to user sessions of a web site. This is shown in FIG. 4. In the window
400, there is a series of rows, such as the rows 402a, 402b and 402c. Each row
10 corresponds to a different type of web page on the site(s) that can be visited by the user.
This is also synonymously referred to herein as a different part of the site(s) that can be
visited by the user. It is noted that only three rows are shown in the window 400 for sake
of illustrative clarity. The rows are divided into a number of columns, such as the
columns 404a, 404b, 404c, 404d and 404e. Each row has a label corresponding to a type
15 of web page in the first column 404a. Thus, the row 402a has the label news,
encompassing news-related web pages of the site(s), the row 402b has the label ent, for
entertainment, encompassing entertainment-related web pages of the site(s), the row 402c
has the label money, encompassing money- or finance-related web pages of the site(s),
etc.

20 Each row has in the columns 404b, 404c, 404d and 404e one or more visible
units, such as blocks, that indicate a value by their shading intensity. In the diagram of
FIG. 4, however, the shading intensity is indicated as a number from 0 to 5 within a
block, where 0 indicates dark, and 5 indicates light. Thus, the block 406 has an intensity

level 5, the block 408 has an intensity level 0, the block 410 has an intensity level 0, and the block 412 has an intensity level 0, etc. This nomenclature is for illustrative clarity only, however, and the invention is not so limited. Furthermore, the number of different intensity levels is also not limited by the invention.

5 Each row has in the column 404b a visible unit that has an intensity level indicating the frequency with which users visit that part of the site, or type of web page on the site, represented by the row. In one embodiment, this is the probability of visiting that type of web page on the site conditioned on inclusion in the cluster represented by the window, or, $p(\text{part}|\text{cluster})$. For example, the block 406 of the row 402a has the
10 maximum intensity level of 5, indicating that this probability is high. Each row also has in the column 404c a visible unit that has an intensity level indicating the frequency with which a user's first visit is to the part of the site represented by the row. For example, the block 408 of the row 402a has the minimum intensity level of 0, indicating that users within this cluster are not likely to visit this part of the site – that is, this type of web page
15 of the site – first.

 Furthermore, each row has in the column 404d a number of visible units that have intensity levels indicating the transition probabilities of the first-order Markov model, such as $p(\text{part } i \text{ at time } t+1, \text{ part } j \text{ at time } t)$, etc., as can be appreciated by those of ordinary skill within the art. Alternatively, the conditional transition probabilities, $p(\text{part } i \text{ at time } t+1 | \text{ part } j \text{ at time } t)$ can be indicated by the visible units within the column
20 404d. Finally, each row has in the column 404e a visible unit that has an intensity level indicating the probability at which the user leaves the site from that part of the site, or type of web page on the site, represented by the row. For example, the block 412 of the

row 402a has the minimum intensity level of 0, indicating that users within the cluster represented by the window 400 are unlikely to leave the web site from the part of the site represented by the row 402a.

Methods and Systems

5 In this section of the detailed description, methods and systems according to varying embodiments of the invention are described. In some embodiments, the methods in particular are computer-implemented. The computer-implemented methods can be realized at least in part as one or more programs running on a computer – that is, as a program executed from a computer-readable medium such as a memory by a processor of
10 a computer, such as the computer shown in and described in conjunction with FIG. 1. The programs are desirably storable on a machine-readable medium such as a floppy disk or a CD-ROM, for distribution and installation and execution on another computer.

Referring to FIG. 5, a flowchart of a method 500 according to an embodiment of the invention is shown. In 502, users are partitioned into clusters, based on similar web
15 browsing behavior thereof, as has been described in the previous section of the detailed description. In 504, windows are displayed on a display device that correspond to the clusters, in a manner consistent with as has also been described in the previous section of the detailed description.

Referring next to FIG. 6, a diagram of a system 600 according to an embodiment
20 of the invention is shown. The system 600 can in one embodiment be implemented within the operating environment described previously in the detailed description in conjunction with FIG. 1. The system 600 includes a display device 602 and a computer program 604. The device 602 can be a computer monitor, such as a flat-panel display

(FPD) or a cathode-ray tube (CRT) display, as known within the art, in one embodiment.

The program 604 is designed to display windows on the device 602 corresponding to clusters into which users have been partitioned based on similar browsing behavior of the users, as has been described in the previous section of the detailed description. In one

5 embodiment, the program 604 can be considered the means for accomplishing this functionality. Furthermore, the program 604 can be executed from a computer-readable medium, such as a hard disk drive or a memory, by a processor in one embodiment.

Conclusion

Although specific embodiments have been illustrated and described herein, it will
10 be appreciated by those of ordinary skill in the art that any arrangement which is calculated to achieve the same purpose may be substituted for the specific embodiments shown. This application is intended to cover any adaptations or variations of the present invention. Therefore, it is manifestly intended that this invention be limited only by the claims and equivalents thereof.

15